第37卷 第14期 计 算 机 工 程 2011年7月
Vol.37 No.14 Computer Engineering July 2011

·软件技术与数据库· 文章编号：1000—3428(2011)14—0053—03 文献标识码：A 中图分类号：TP311

基于控制流挖掘的知识维护算法

张 嬰[1]，朱东华[1]，李 明[2]

(1. 北京理工大学管理与经济学院，北京 100081；2. 北京航空航天大学经济管理学院，北京 100191)

摘 要：针对知识库中知识维护的控制问题，在分析过程挖掘特点的基础上，提出一个基于过程挖掘的知识维护过程框架和一种基于控制流挖掘的知识维护算法，用于对知识维护进行优化，保证知识库中知识的质量，提高知识重用和共享效果。实验结果证明，该算法能较好地反映知识维护的实际情况。

关键词：知识管理；控制流挖掘；知识库；知识维护；过程挖掘

Knowledge Maintenance Algorithm Based on Control Flow Mining

ZHANG Zhao[1], ZHU Dong-hua[1], LI Ming[2]

(1. School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China;
2. School of Economics and Management, Beihang University, Beijing 100191, China)

【Abstract】Aiming at the problem of controlling knowledge maintenance process, a framework of knowledge maintenance process and a novel knowledge maintenance algorithm based on control flow mining are proposed according to the characteristics of process mining. The process of knowledge maintenance is optimized by the algorithm, which ensures the quality of knowledge in the knowledge repository and improves the effects of knowledge reusing and sharing. Test result can reflect the actual situation of the knowledge maintenance well.

【Key words】knowledge management; control flow mining; knowledge repository; knowledge maintenance; process mining

DOI: 10.3969/j.issn.1000-3428.2011.14.016

1 概述

知识的质量直接影响了知识共享的效果[1]。低质量的知识会降低知识共享的效果，错误的知识甚至可能导致严重错误的发生。因此，在共享知识之前需要对知识进行一系列检查以确保知识的高质量[2]。大多数企业的知识维护系统采用工作流系统支撑知识维护过程，由此可以得到大量的知识维护过程日志。这些日志记录了每个知识维护过程实例的执行情况。因此，本文采用过程挖掘技术对知识维护日志进行分析以获得实际的知识维护过程模型。

2 过程挖掘

过程挖掘是利用已有的过程信息，使用过程挖掘技术获得过程模型，它解决了预先定义的模型与实际模型不一致的问题，对充分理解现有过程执行情况有很大的帮助[3]。目前主要从2个角度对任务执行日志进行挖掘：控制流挖掘和组织挖掘。控制流挖掘是通过分析日志中任务执行的序列发现任务之间的依赖关系[4]。组织挖掘是将任务执行者进行聚类，并且发现聚类之间的关系。本文主要从控制流挖掘的角度研究知识维护算法。

3 基于控制流挖掘的知识维护

在知识维护过程挖掘中，无效的任务要被删除以提高知识维护的效率，同时部分新增的任务将被增加到知识维护过程模型中，以更加全面地对知识进行检查。由于知识维护任务只有被合格的任务执行者执行才能起到应有的检查效果，因此找到具有相同知识能力的任务执行者是知识维护过程挖掘又一重要任务。

基于过程挖掘的知识维护过程如下：首先，知识按照预定义的原始过程模型不断地进入和离开知识库，在这个过程中产生的日志被保存在知识维护日志库中，当积累一段时间后，对这些日志进行预处理，然后进行控制流挖掘，最终将挖掘的结果提供给管理员，作为优化知识维护过程的重要参考。

3.1 预定义的控制流模型与案例数据

知识维护过程是由知识维护任务构成的。每个知识库具有相应的知识维护过程。比如，显性知识的获取过程由审查和审批2个知识维护任务构成，而经验总结类知识(隐性知识)的获取过程由审查、审核、审定和审批4个知识维护任务所构成。预先定义的任务之间的依赖关系即原始的控制流模型如图1所示。

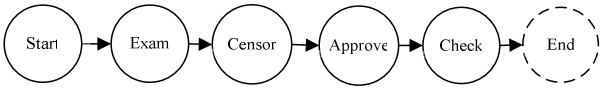


图1 预定义的经验总结类知识获取控制流模型

图1中的节点表示知识维护任务，边表示维护任务之间的依赖关系。这些知识维护任务的含义如下：(1)Start表示开始的任务，即初始化知识维护过程；(2)Exam表示审查任务，是对知识进行最基本的检查；(3)Censor 表示对知识进行审核；(4)Approve 表示对知识进行审定；(5)Check 表示对知识的确认及审批；(6)End表示任务的结束。

3.2 基于控制流挖掘的知识维护算法

本文主要研究任务以及任务之间的依赖关系，因此，需要将日志进行预处理，删除日志中的执行者信息，预处理后

基金项目：国家自然科学基金资助项目(70639001)；中国博士后科学基金资助项目(20090450297)

作者简介：张 嬰(1974－)，女，博士，主研方向：知识管理，知识系统工程；朱东华，教授；李 明，博士

收稿日期：2010-12-05 **E-mail：**zhaoer1974@163.com

的日志信息如表 1 所示。其将作为本文的案例数据，对所提出的算法进行解释。

表 1 处理日志 R 后的日志 *L*

ID	日志
1	Start, Exam, Start, Exam, Secrecy, Censor, Start, Exam, Secrecy, Censor, Approve, Check
2	Start, Exam, Secrecy, Start, Exam, Secrecy, Censor, Approve, Check
3	Start, Exam, Secrecy, Start, Exam, Secrecy, Censor, Approve, Start, Exam, Secrecy, Censor, Approve, Check
4	Start, Exam, Secrecy, Censor, Approve, Start, Exam, Secrecy, Censor, Approve, Check
5	Start, Exam, Censor, Secrecy, Start, Exam, Censor, Secrecy, Approve, Check
6	Start, Exam, Censor, Secrecy, Approve, Check

知识维护过程可以由图 $G = (V, E)$表示，其中，V 表示知识维护任务的集合；E 表示任务之间的依赖关系。E 的初始值为 $\varnothing$，V 的初始值为开始任务 S。L 中的每一个知识维护过程执行实例 p 是一个以任务 S 为开始的任务序列，$p=\{s,t_1,t_2,\cdots,t_n\}$。基于控制流挖掘的知识维护算法步骤如下：

步骤 1 对于 L 中的每一个 p，p 中除 S 以外的每一个任务 t，如果 t 的输出为负，并且 V 包含 t，那么 t 的计数加 1，如果 t 的输出为负，但 t 在 V 中不存在，则将 t 加入到 V 中。

步骤 2 将 V 中除任务 S 外计数小于给定阈值 δ 的任务删除。

步骤 3 p’是重复的任务被删除后的过程实例。L’是 p’的集合。对于 L 中的每一个 p，最后一个以任务 S 开始的子序列被连接到 p’中。处理完 p 后，将 p’加入到 L’中。

步骤 4 对于 L’中每一个处理过的过程实例 p’，如果其中的任务 u 出现在 v 之前，则将边(u, v)加入到 E 中。

步骤 5 删除 E 中存在的双向边。

步骤 6 对于 G 中每一个强连通部分，删除 E 中强连通部分的任务之间的所有边。

步骤 7 对 L’中每一个处理过的过程实例 p’：

(1)找到 G 的生成子图。

(2)计算子图的传递规约。

(3)标记 E 中传递规约的边。

步骤 8 删除 E 中没有标记的边。

步骤 9 返回(V, E)。

知识维护过程挖掘算法分为 3 个部分。第 1 部分包括步骤 1 和步骤 2，用于将噪音任务和无效任务删除。第 2 部分包括步骤 3，用于将日志中重复的子过程清除。第 3 部分包括步骤 4~步骤 9，用于发现保留下来的任务之间的依赖关系。

下面结合表 1 中的日志 L 对上述算法进行说明。

步骤 1 删除无效的任务。如果某任务的所有执行结果都是通过，说明知识在这个方面不会出现错误，不必进行检查，该任务是无效任务。

表 2 是经步骤 1 后得到的结果。从中可以看出，由于 Check 的有效执行次数为 0，即 Check 所有执行结果都是通过，因此其被认为是无效的，需要被删除。

表 2 有效的任务

任务	数量
Start	13
Exam	13
Secrecy	12
Censor	13
Approve	8
Check	0

步骤 2 删除日志中的噪声任务。人工添加的任务中出现次数小于给定阈值的任务被认为是噪声任务。在本例中，设定阈值 δ=2，由于人工添加的任务出现次数均大于阈值，因此没有任务被删除。

步骤 3 处理过程实例中的重复子过程，最后一个以 Start 开头的子过程被保留下来。经过步骤 3 后，获得 L’：

L’={(Start, Exam, Secrecy, Censor, Approve), (Start, Exam, Censor, Secrecy, Approve)}

步骤 4 定义任务之间的依赖关系。

图 2 是经过步骤 4 后得到的结果。

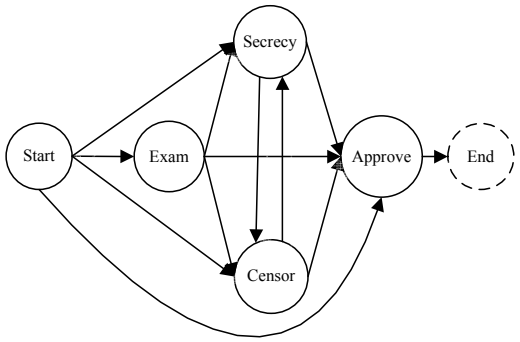


图 2 日志 *L* 中的任务依赖关系

步骤 5、步骤 6 识别并行的任务，这些任务的执行顺序可以发生变化。

图 3 是经过步骤 5 和步骤 6 后的挖掘结果。其中，由于 Secrecy 可以在任务 Censor 之前执行，而 Censor 也可以在 Secrecy 之前执行，说明 Secrecy 和 Censor 之间不存在依赖关系，是并行的，因此 Censor 和 Secrecy 之间的依赖关系需要被删除。

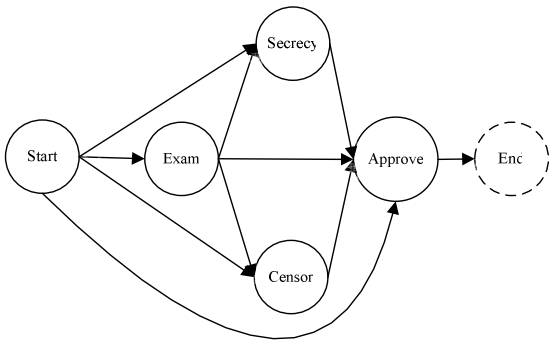


图 3 日志 *L* 的最小共形图

步骤 7、步骤 8 保留任务之间实际存在的依赖关系。

图 4 是经过步骤 7 和步骤 8 得到的结果。为了能够发现并行的任务，在步骤 4 中加入了直接依赖关系，而由于在实际的知识维护过程中这些依赖关系并不存在，因此需要在模型中将其删除。比如在图 4 中，Start 和 Secrecy 之间的依赖关系由于没有在任何过程实例中出现，因此被删除掉，得到了最终的挖掘结果。

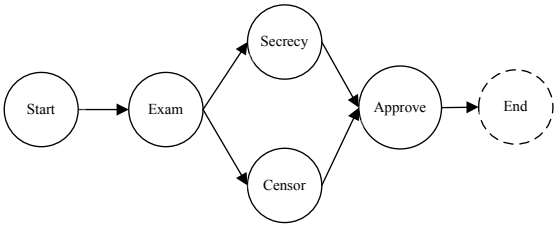


图 4 挖掘出的控制流模型

3.3　知识维护过程优化

在完成对知识维护日志的控制流挖掘后，可以参考挖掘的结果对原始的过程模型进行优化，使系统中预定义的原始模型符合实际的知识维护需求。

对比图 4 的控制流挖掘结果与图 1 的原始过程模型可以看出，Check 被认为是无效任务而被删除，Secrecy 被增加到控制流模型中，Secrecy 与 Censor 是并行关系。根据以上结果可以从以下 2 点对系统中的模型进行优化：

(1)Check 任务的执行者一般是参考前面的 Approve 任务执行者的意见，经过一段时间的运行，通过挖掘结果可以看出，Check 任务的执行结果均为通过，说明该任务无法起到应有的检查效果，与前面的 Approve 任务重复，所以，需要将 Check 任务从系统预定义的知识维护过程模型中删除。

(2)Secrecy 是为了满足企业保密需求而由各单位自行添加的知识维护任务。由于目前企业尚未对 Secrecy 的执行情况进行统一，因此得到的结果是并行的。由此可以看出，Secrecy 与 Censor 实际上没有严格的先后顺序，在对知识维护过程模型进行优化时，可以同时将 Secrecy 与 Censor 设置为并行关系。

4　实验验证

4.1　评价指标

实验目的是检验经过控制挖掘出的知识维护模型能否真实反映出日志中的知识维护过程。本文采用查全率和查准率这 2 个评价指标评价。

4.2　实验结果

实验所使用的数据集是从某企业知识维护系统中收集的知识维护日志，并且只采用已经结束的过程实例。处理后的数据集中包含了 6 000 条过程实例。实验结果如表 3 和图 5 所示。从表中可以看出，查准率和查全率随挖掘中所使用数据集实例的比例的提高而提高。在刚开始时，所使用的实例数量较少，大多数节点被认为是无效的，获得的依赖关系也少，因此，实验结果较差。而随着实例数量的增加，更多的节点被保留，同时更多的依赖关系被获得，算法性能在提高。由图 5 可以看出，当挖掘时使用的实例数量超过数据集中实例数量的 40%时，查准率和查全率分别为 89.9%和 82.1%。当使用的比例超过 60%时，查全率和查准率分别为 92.4%和 87.6%。总体上，本文的控制流挖掘算法能够较好地反映出实际的知识维护实际情况。

表 3　控制流挖掘实验结果

数据集中实例的比例/(%)	查准率	查全率
10	0.301	0.232
20	0.589	0.401
30	0.831	0.659
40	0.899	0.821
50	0.913	0.839
60	0.924	0.876
70	0.938	0.893
80	0.948	0.917
90	0.959	0.918
100	0.976	0.931

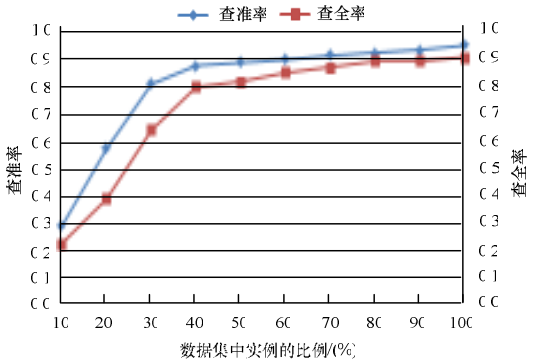


图 5　算法查准率与查全率曲线

5　结束语

本文主要研究了企业中的知识维护过程挖掘算法。从控制流的视角对企业知识维护过程日志进行分析。通过实验验证基于控制流挖掘的知识维护算法可以较好地反映出实际知识维护过程的执行情况。下一步将从组织挖掘的角度对知识维护过程做进一步的探讨。

参考文献

[1] Rao Lila, Osei-Bryson K M. Towards Defining Dimensions of Knowledge Systems Quality[J]. Expert Systems with Applications, 2007, 33(2): 368-378.

[2] 刘昌余, 王红军, 陈佳鑫, 等. 面向工作流和服务的电子邮件系统模型[J]. 计算机工程, 2010, 36(15): 68-71.

[3] van der Aalst W M P, Weijters A J. Process Mining: A Research Agenda[J]. Computers in Industry, 2004, 53(3): 231-244.

[4] van der Aalst W M P, van Dongen B F, Herbst J, et al. Workflow Mining: A Survey of Issues and Approaches[J]. Data and Knowledge Engineering, 2003, 47(2): 237-267.

编辑　张　帆

(上接第 52 页)

参考文献

[1] Daylight Chemical Information Systems, Inc.. Daylight Theory Manual-daylight Version 4.9[EB/OL]. (2008-02-01). http://www.daylight.com/dayhtml/doc/theory/index.html.

[2] Milo T, Suciu D. Index Structures for Path Expressions[C]// Proceedings of the 7th International Conference on Database Theory. Jerusalem, Palestine: [s. n.], 1999: 277-295.

[3] Goldman R, Widom J. DataGuides: Enabling Query Formulation and Optimization in Semistructured Database[C]//Proceedings of the 23rd International Conference on Very Large Data Bases. Athens, Greece: [s. n.], 1997: 436-445.

[4] Shasha D, Wang J T L, Giugno R. Algorithmics and Applications of Tree and Graph Searching[C]//Proceedings of the 21st Symposium on Principles of Database Systems. Los Angeles, California, USA: ACM Press, 2002: 39-52.

[5] Yan Xifeng, Yu P S, Han Jiawei. Graph Indexing: A Frequent Structure-based Approach[C]//Proceedings of International Conference on Management of Data. Paris, France: [s. n.], 2004: 335-346.

[6] 黄崇本, 陶剑文. 一种新颖的对比子图索引算法[J]. 计算机工程, 2009, 35(5): 64-67, 96.

编辑　陆燕菲